Text Mining and Sentiment Analysis of Tourist Reviews for Heritage Attractions in Anuradhapura, Sri Lanka

Hashini T. Wickremasinghe[1]*

[1] *Deakin University, Australia*

Abstract

In today's digital era, online reviews and user-generated content play a crucial role in shaping travel decisions globally, offering rich yet underutilized insights into visitor experiences. However, these valuable insights into visitors' experiences remain largely untapped and underutilized. This study investigates the application of text mining and machine learning–based sentiment classification to analyze tourist reviews of Anuradhapura's heritage sites in Sri Lanka, aiming to provide data-driven insights for tourism management. Reviews from Google and TripAdvisor, spanning 2018–2024, were preprocessed and classified into positive, neutral, and negative sentiments using a decision tree model. The model achieved an overall accuracy of 80.85% and substantial agreement (kappa = 0.634), effectively capturing dominant positive and negative sentiments. Analysis revealed that positive reviews were influenced by aesthetic appeal, architectural significance, and spiritual engagement, while negative sentiments reflected operational challenges, unmet expectations, and underwhelming site conditions. The study underscores the potential of automated sentiment analysis to guide heritage site management, inform strategic interventions, and enhance visitor experiences, offering a scalable methodology adaptable to other heritage destinations globally.

Keywords: Text Mining, Sentiment Analysis, Natural Language Processing, Machine Learning, Tourism Analytics, Sri Lanka

Introduction

In the modern digital landscape, online reviews and user-generated content have become powerful forces shaping travel choices across the globe, offering abundant but often untapped insights into visitors' experiences. Platforms like TripAdvisor, Google Reviews, and social media host a rapidly growing volume of traveller feedback, offering valuable experiential data about tourism destinations (Mutalib et al., 2021). However, much of this data is unstructured and conveyed in natural language, making it challenging to analyze using traditional methods. Conventional approaches such as surveys or field interviews, are constrained by small sample sizes, resource-intensive, limited in scope, single-dimensional data, and subjective biases, making it difficult to comprehensively capture visitors' fine-grained perceptions in diverse scenarios and often fail to capture the dynamic, real-time sentiments expressed by tourists online (Yuan et al., 2025). Consequently, tourism managers

and policymakers lack comprehensive, timely insights necessary for effective decision-making.

*Corresponding Author- *s224932797@deakin.edu.au*| *hashinitw@gmail.com*

Tourism is an important part of Sri Lanka's economy, especially in the heritage tourism sector. Despite the cultural and economic significance of sites like Anuradhapura, the country's ancient UNESCO World Heritage city, modern data analytics have been used only minimally to understand visitor perceptions systematically. Harnessing advanced computational methods, particularly sentiment classification within the field of Natural Language Processing (NLP), offers the potential to transform vast amounts of online review data into actionable insights. By automatically identifying positive, neutral, or negative sentiments, such techniques enable a deeper understanding of visitor experiences and satisfaction drivers at scale.

This study aims to develop a machine learning framework capable of efficiently classifying tourist sentiments from online reviews related to Anuradhapura's attractions. The goal is to uncover key terms and areas for improvement that may otherwise go unnoticed in traditional analysis. The findings will provide tourism authorities, local policymakers, destination managers, and digital marketers with data-driven guidance to enhance service quality, preserve cultural heritage, and optimize visitor experiences.

Objective

To develop and evaluate text mining–based sentiment classification models on tourist reviews of Anuradhapura's heritage sites for data-driven tourism management.

Literature Review

Sentiment analysis (SA) has emerged as a pivotal tool in tourism research for capturing visitor perceptions and guiding evidence-based management. SA, or opinion mining, involves extracting attitudes, emotions, and evaluations from textual data, enabling a deeper understanding of how tourists assess destinations, services, and experiences (Xiang, Schwartz, Gerdes & Uysal, 2015). In tourism, these analyses provide critical insights for destination marketing, experience enhancement, and resource allocation (Surugiu et al., 2023).

Recent studies demonstrate the growing reliance on natural language processing (NLP) and machine learning methods for sentiment classification. Sentiment analysis is often operationalized as a supervised learning task, training models on labeled datasets to predict the polarity of new texts. Transformer-based models, such as BERT and RoBERTa, have enhanced contextual comprehension and classification performance, outperforming traditional lexicon- or feature-based approaches (Viñan-Ludeña & de Campos, 2022). Applications in heritage tourism illustrate the potential of these methods: Gulati (2022) analyzed Twitter sentiments for Indian heritage sites, while Yuan et al. (2025) examined architectural heritage perceptions to inform sustainable conservation. Studies by Singgalen (2023) and Mutalib et al. (2021) further highlight the utility of machine learning models, including decision trees, support vector machines, and regression models, for predicting visitor behavior and classifying sentiments.

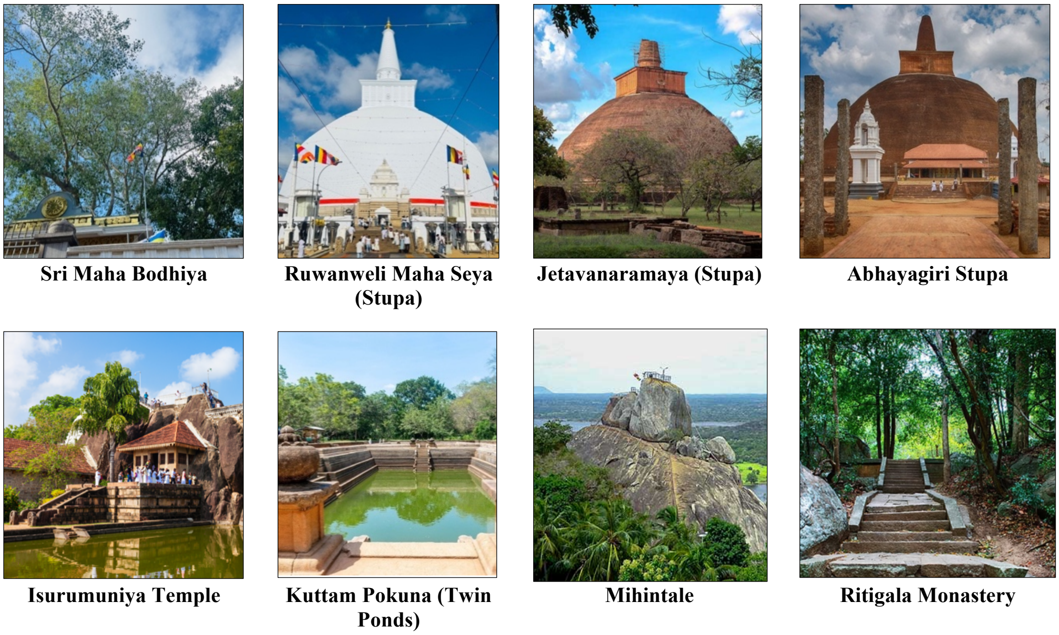
Despite these advancements, research in South Asia remains limited. In Sri Lanka, heritage sites such as Anuradhapura are globally significant, yet there is scant work applying text mining–based sentiment classification to capture tourist feedback. The linguistic, cultural, and contextual nuances of reviews in this region remain underexplored. This study addresses this gap by developing and evaluating sentiment classification models on foreign visitor reviews of Anuradhapura’s heritage sites, providing actionable insights for data-driven tourism management and strategic heritage preservation.

Methods

This study analyzed tourist reviews of major heritage sites in Anuradhapura to develop sentiment classification models for data-driven tourism management. The sites included Ruwanweli Maha seya, Sri Maha Bodhiya, Jetavanaramaya, Isurumuniya Temple, Kuttam Pokuna (Twin Ponds), Mihintale, Abhayagiri Stupa and Ritigala Monastery. These were ranked as top visited places in Anuradhapura by both platforms of Google Reviews and TripAdvisor.

Figures 1-7

Selected Heritage Sites



Sri Maha Bodhiya

Ruwanweli Maha Seya (Stupa)

Jetavanaramaya (Stupa)

Abhayagiri Stupa

Isurumuniya Temple

Kuttam Pokuna (Twin Ponds)

Mihintale

Ritigala Monastery

Source: Author (2025), Tripadvisor (2025), Story of Sri Lanka (2024), Punchihewa (2021), Travel Setu (2025)

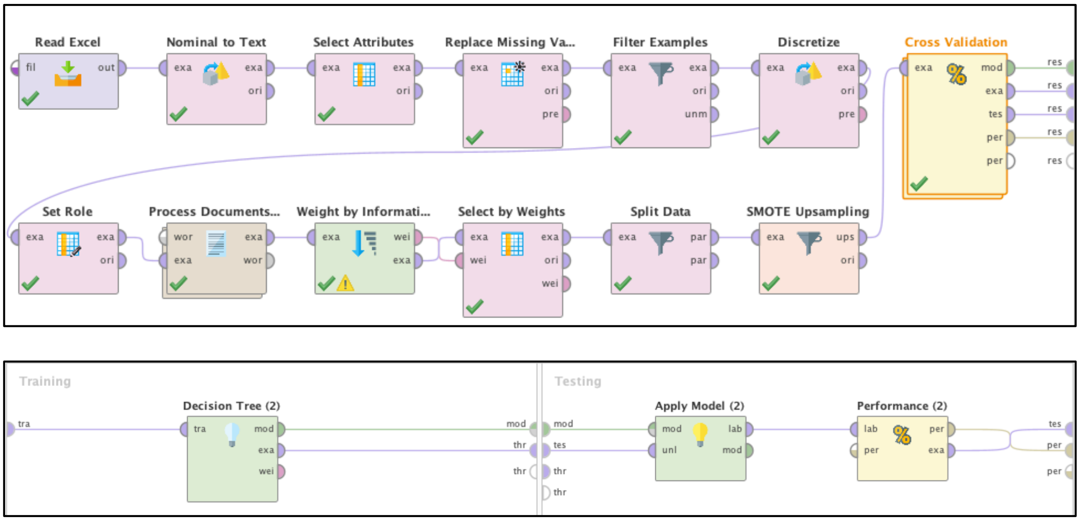
Reviews were collected from public platforms, specifically Google Reviews and TripAdvisor during January 2018 to December 2024 period, using the Instant Data Scraper tool. The reviews from foreigner visitors only were retained, and local visitors or irrelevant content was removed to ensure dataset quality and consistency.

The collected dataset underwent a comprehensive text mining process using the Altair AI Studio tool, to convert unstructured textual data into structured features suitable for modeling. Preprocessing included transforming text to lowercase, tokenization, stemming, stopword removal, filtering tokens by length, and noise elimination such as punctuation, numbers, and special characters. After preprocessing, a weigh-select dimensionality reduction method was applied to retain the most informative features.

To create sentiment labels for supervised learning, the star ratings provided by reviewers were discretized based on upper limits: Positive (≤5.0), Neutral (≤3.0), and Negative (≤2.0), allowing textual content to be mapped consistently to sentiment categories. The dataset was then split into training and testing sets using a 70:30 ratio. To address class imbalance, SMOTE upsampling was applied to the training set, ensuring that minority classes were adequately represented.

Figure 8

Workflow for Sentiment Classification of Tourism Reviews



Source: Author (2025)

A decision tree classifier was trained on the processed training set using k-fold cross-validation to evaluate performance, improve robustness, and prevent overfitting. Model outputs were interpreted using term importance rankings and performance metrics, linking key terms such as “beauty” and “beautifully” with positive sentiment, and “nothing,” “disappoint,” and “remove” with negative sentiment.

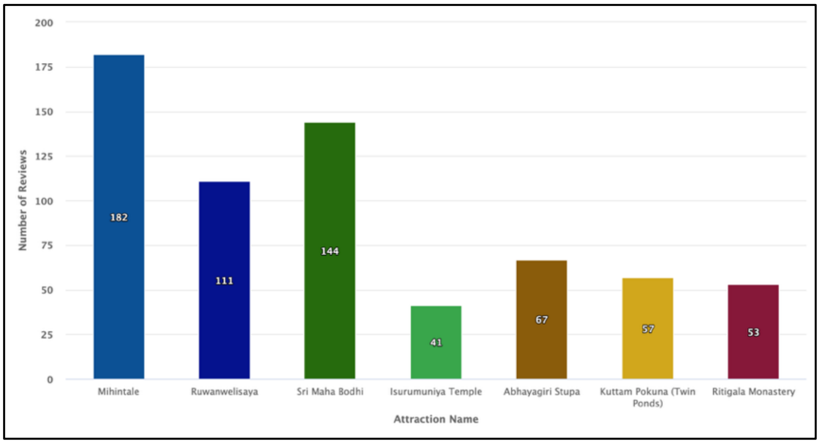
This methodology integrates systematic text preprocessing, dimensionality reduction, sentiment discretization, upsampling for class balance, and rigorous model evaluation to provide a structured, data-driven framework for analyzing tourist sentiment and supporting heritage tourism management strategies in Anuradhapura.

Results and Discussion

The chart shown in figure 09 indicates that Mihinthale received the highest number of reviews (182), followed by Sri Maha Bodhi (144), while Isurumuniya Temple had the fewest reviews (41).

Figure 09

Number of User Reviews of Each Location

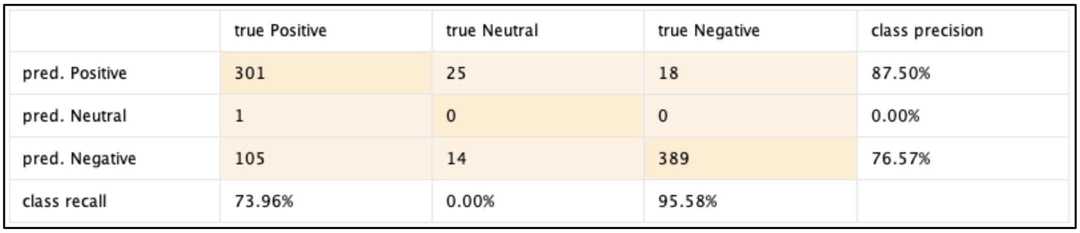


Source: TripAdvisor and Google Reviews (2018–2025)

Model Performance

The sentiment classification model, based on a decision tree algorithm, achieved an overall accuracy of 80.85% with a kappa of 0.634, indicating substantial agreement beyond chance.

Figure 10
Confusion Matrix of Sentiment Classification



	true Positive	true Neutral	true Negative	class precision
pred. Positive	301	25	18	87.50%
pred. Neutral	1	0	0	0.00%
pred. Negative	105	14	389	76.57%
class recall	73.96%	0.00%	95.58%	

Source: Author (2025)

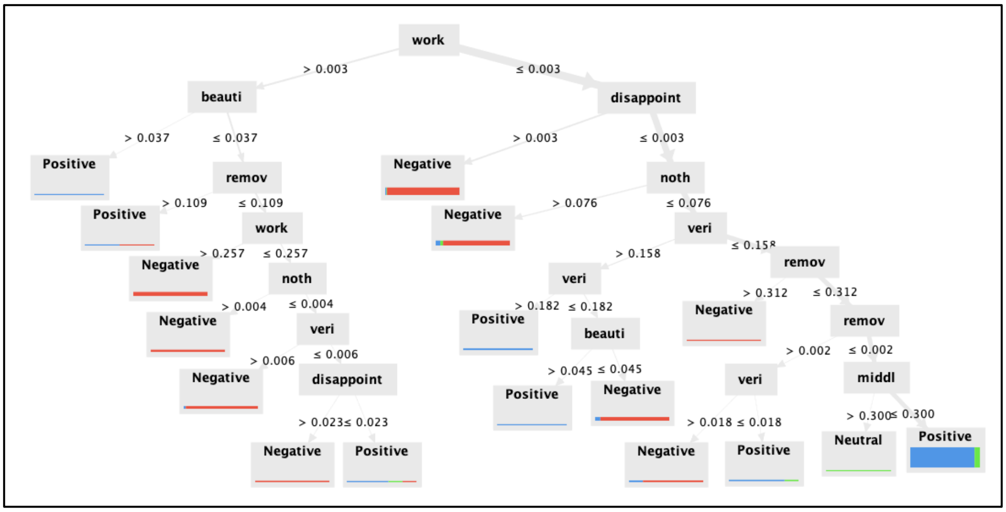
It performed strongly in identifying positive (301 correct) and negative (389 correct) reviews, while neutral reviews were rarely captured. Weighted mean recall (56.5%) and precision (57.5%) were moderate, reflecting challenges in distinguishing minority classes. Overall, the model effectively captured the dominant positive and negative sentiment patterns in tourist reviews.

Sentiment Classification

The sentiment analysis of tourist reviews from Anuradhapura's heritage sites, conducted using a text mining approach with a decision tree–based classification model, revealed clear patterns in how specific words and phrases influence positive or negative evaluations. Consistent with prior research in heritage tourism sentiment analysis (Singgalen, 2023), key terms emerged as strong predictors of visitor sentiment.

Figure 11

Decision Tree of Sentiment Classification



Source: Author (2025)

Words associated with aesthetics and admiration, such as "beauty" and "beautifully," were closely linked to positive sentiment. For instance, one visitor noted, "*The grounds are beautifully maintained, and there's a sacred atmosphere that invites you to pause and reflect*," highlighting how both visual appeal and emotional ambiance contributed to a highly favourable assessment. Another review emphasized architectural and scenic value, describing the brick stupa as "*a real Buddhist reliquary work… a beautiful walk in the park with some remains*," illustrating how references to architectural achievement and the surrounding environment reinforce positive perceptions. The results of this study are consistent with the findings of Yuan et al. (2025), who reported that positive visitor sentiments are strongly associated with aesthetic appeal and cultural significance of heritage sites.

Negative sentiment was most strongly associated with terms reflecting unmet expectations or dissatisfaction. The word "disappoint" emerged as a major indicator, appearing in reviews such as, "*In my opinion, in general, the whole site is pretty disappointing. As the country has better attractions (i.e., Polonnaruwa), I would visit Anuradhapura only in case you have some spare time*," and was linked to 130 negative cases in the decision tree, highlighting its robust predictive power. Similarly, "nothing" appeared in reviews expressing underwhelming experiences, including, "*We came here and it was just another stupa! Nothing special to make your way here if you don't have time… There is nothing outstanding,*" with 72 negative cases captured in the tree. The term "work" was another significant predictor of negative sentiment, appearing in 63 cases, reflecting visitors' perception of effort or inconvenience in exploring certain sites. Secondary terms such as "remove" and "very" also appeared in the tree. "Remove" was associated with operational discomfort, particularly the need to remove shoes or hats or parasols and was linked to 13 negative cases and several positive ones depending on context, as reflected in the reviews: "*No shoes allowed in the complex, as usual, but this gets extremely uncomfortable when the weather is hot,*" and "*We arrived at the hottest part of the day to be told that we could not wear hats or even parasols… Very unpleasant experience. Avoid.*" These patterns mirror findings in other heritage tourism studies, such as Viñan-Ludeña and de Campos (2022), who identified operational inefficiencies and service delays as key drivers of visitor dissatisfaction, and Yuan et al.

(2025), who found that site management and environmental factors, including safety and comfort, were major contributors to negative visitor perceptions.

"Very," functioning primarily as an intensifier, contributed to sentiment classification in 35 negative cases and 16 positive cases when combined with other terms, demonstrating that intensifiers can subtly modulate perceived sentiment. The term "middl" captured neutral sentiment in a small subset of cases (32), indicating reviews that were neither strongly positive nor negative.

Overall, the findings emphasize the multidimensional nature of tourist experiences at heritage sites. Positive reviews were primarily shaped by aesthetic appeal, architectural significance, and spiritual engagement, whereas negative reviews reflected unmet expectations, operational challenges, and underwhelming site conditions. Decision tree–based sentiment classification effectively captured these patterns, providing nuanced insights into how specific textual cues signal broader visitor experiences. Such insights can guide heritage tourism managers in enhancing visitor satisfaction by emphasizing admired features while addressing sources of dissatisfaction.

Policy Level Implications

The study demonstrates the transformative potential of sentiment classification in tourism analytics, especially for heritage-rich destinations like Anuradhapura. By systematically analyzing tourist reviews, it provides data-driven insights to enhance visitor experiences, inform preservation efforts, and support strategic destination management. To enhance visitor experiences at Anuradhapura's heritage sites, several targeted strategies are recommended. Visitor comfort can be improved by providing shaded rest areas, water points, seating along long or steep routes, and solutions such as mats for sites with mandatory shoe- or hat-removal rules. Interpretation and educational support should include multilingual signage, explanatory boards, and digital guides or apps with audio-visual storytelling to highlight historical and cultural significance. Regular maintenance of ruins, ponds, and monuments is essential to preserve visual appeal and minimize disruptions during restoration. Entry fees should reflect the value of the experience and be clearly communicated. Finally, a sentiment dashboard can enable real-time monitoring of online reviews, guiding resource allocation, operational policies, and marketing strategies based on visitor feedback.

Limitations and Future Research Directions

This study has several limitations. First, the analysis relies primarily on online reviews and user-generated content, which may be biased toward more vocal or digitally active visitors, potentially underrepresenting other visitor segments. Second, the research focuses on specific heritage sites, limiting the generalizability of the findings to other locations or types of tourism properties. Third, the sentiment analysis is based on textual data and may not fully capture nuanced visitor experiences, such as emotions or contextual factors influencing satisfaction. Further, contextual ambiguities such as sarcasm or mixed sentiments, may also complicate accurate classification. Finally, external factors like seasonal variations, special events, or temporary disruptions were not extensively controlled for, which may influence visitor perceptions.

Future studies could incorporate mixed methods, such as surveys, interviews, or observational studies, and complement textual analysis and provide a deeper understanding of visitor experiences. Research could also explore the impact of targeted interventions, such as infrastructure improvements or digital interpretation tools, on visitor satisfaction over time. Finally, future studies should also explore aspect-based sentiment analysis to capture opinions on specific attributes like cleanliness and accessibility, integrate spatio-temporal data to visualize sentiment trends over time and location, profile tourist personas by combining sentiment with demographic data, and design hybrid recommender systems that incorporate sentiment insights for personalized travel planning.

Conclusion

This study demonstrates the value of applying text mining and decision tree–based sentiment classification to analyze tourist reviews of Anuradhapura's heritage sites. The findings reveal clear patterns in visitor perceptions, highlighting both strengths, such as aesthetic appeal and cultural significance, and areas needing improvement, including operational challenges and unmet expectations. By providing actionable, data-driven insights, this approach enables heritage managers, policymakers, and marketers to enhance visitor experiences, preserve cultural assets, and implement targeted interventions. The methodology offers a scalable, adaptable framework for heritage tourism analytics in Sri Lanka and beyond, supporting evidence-based, sustainable tourism development.

Acknowledgement

The author gratefully acknowledges Professor Rens Scheepers, Department of Information Systems and Business Analytics, School of Business and Law, Deakin University, Australia, for his guidance and insights, which provided valuable understanding of text mining and sentiment classification applied in this study.

References

Gulati, S. (2022). Tapping public sentiments on Twitter for tourism insights: A study of famous Indian heritage sites. *International Hospitality Review, 36*(2), 244–257. https://doi.org/10.1108/IHR-03-2021-0021

Mutalib, S., Razali, A. H., Kamarudin, S. N. K., Halim, S. A., & Abdul-Rahman, S. (2021). Prediction of tourist visit in Taman Negara Pahang, Malaysia using regression models. *International Journal of Advanced Computer Science and Applications, 12*(12), 746–754. https://doi.org/10.14569/IJACSA.2021.0121292

Punchihewa, S. (2021). *Recognition of Mihintale as a World Heritage Site is long overdue*. [Photograph]. *The Island*. Retrieved from

Singgalen, Y. A. (2023). Analisis Sentimen Top 10 Traveler Ranked Hotel Di Kota Makassar Menggunakan Algoritma Decision Tree Dan Support Vector Machine. *KLIK: Kajian Ilmiah Informatika dan Komputer, 4*(1), 323–332. https://doi.org/10.30865/klik.v4i1.1153

Surugiu, C., Surugiu, M.-R., & Grădinaru, C. (2023). Targeting creativity through sentiment analysis: A survey on Bucharest city tourism. *SAGE Open, 13*(2), 1–17. https://doi.org/10.1177/21582440231167346

Travel Setu. (2025). *Jetavanaramaya Tourism Guide*. [Photograph]. Retrieved from https://travelsetu.com/guide/jetavanaramaya-tourism

Tripadvisor. (2025). *Abhayagiri Dagaba, Anuradhapura*. [Photograph].

Viñan-Ludeña, M. S., & de Campos, L. M. (2022). Discovering a tourism destination with social media data: BERT-based sentiment analysis. *Journal of Hospitality and Tourism Technology, 13*(5), 907–921. https://doi.org/10.1108/JHTT-09-2021-0259

Wonders of Ceylon. (2024). *Isurumuniya Temple, Anuradhapura*. [Photograph]. Retrieved from https://www.wondersofceylon.com/isurumuniya-temple/

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013

Yuan, H., Ke, R., & Xie, X. (2025). Sentiment analysis of visitor perceptions on architectural heritage: A case study of Phoenix Ancient Town for sustainable conservation and development. *Journal of Asian Architecture and Building Engineering*. Advance online publication. https://doi.org/10.1080/13467581.2025.2540079